(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 December 2014 (11.12.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/196971 A1

(51) International Patent Classification:
G10L 17/00 (2013.01)

(21) International Application Number:
PCT/US2013/044413

(22) International Filing Date:
6 June 2013 (06.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **THOMSON LICENSING** [FR/FR]; 1-5 rue Jeanne d'Arc, F-92130 Issy les Moulineaux (FR).

(72) Inventor; and
(71) Applicant : **SCHULTZ, Mark, Alan** [US/US]; 13652 State Highway 14 W., Billings, Missouri 65610 (US).

(74) Agents: **SHEDD, Robert, D.** et al.; c/o Thomson Licensing LLC, 2 Independence Way Suite 200, Princeton, New Jersey 08540 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: VOICE RECOGNITION AND IDENTIFICATION

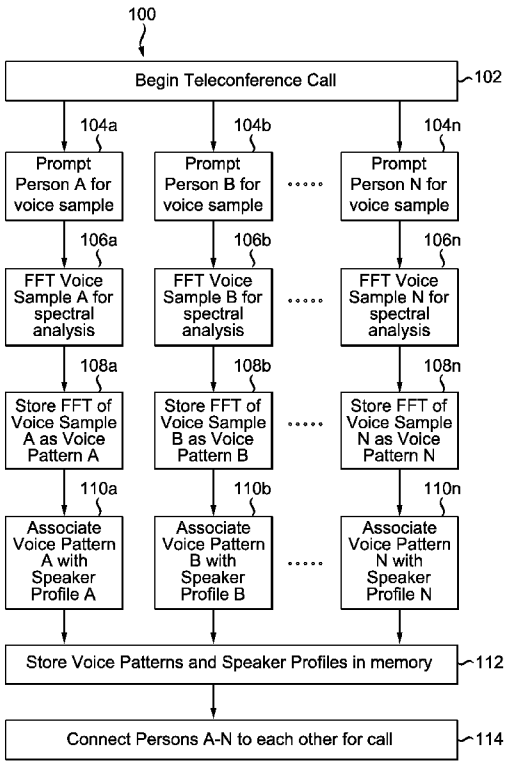


FIG. 1

(57) Abstract: A voice recognition method and system for identifying a person on a display screen based on when the person's voice is recognized is disclosed. The method includes obtaining a speaker profile associated with a speaker, the speaker profile including a signature pattern and identifying information associated with the speaker, the signature pattern having been created from a Fast Fourier Transform (FFT) of a first voice sample of the speaker's voice; receiving a second voice sample; converting the second voice sample into a voice pattern using FFT; cross correlating the voice pattern with the signature pattern, resulting in a cross correlation coefficient R, where $0 \leq R \leq 1$; and sending the identifying information associated with the speaker profile when the cross correlation coefficient R is above a threshold.

WO 2014/196971 A1

Published:

— *with international search report (Art. 21(3))*

VOICE RECOGNITION AND IDENTIFICATION

BACKGROUND

[0001] As teleconferencing and video conferencing increase in usage, knowing who is speaking at any given time is important. Many times, the persons on a teleconferencing call may know each other's names but have never met before and are not familiar with each other's voices. In such circumstances, when one person speaks during a teleconferencing call, the other persons on the call may not know who is speaking without the person self-identifying him or herself during the call. However, not every person on a call may remember to self-identify during the call, which may cause confusion for the other persons on the call and cause them to ask who was speaking, which wastes time.

SUMMARY

[0002] In view of the foregoing, a voice recognition method for identifying a person on a display screen based on when the person's voice is recognized is disclosed. The method can be implemented in a number of systems where such identification is desirable, such as teleconferencing, video conferencing, multiplayer video game platforms, and the like.

[0003] In one embodiment, the method comprises obtaining a speaker profile associated with a speaker, the speaker profile including a signature pattern and identifying information associated with the speaker, the signature pattern having been created from a Fast Fourier Transform (FFT) of a first voice sample of the speaker's voice; receiving a second voice sample; converting the second voice sample into a voice pattern using FFT; cross correlating the voice pattern with the signature pattern, resulting in a cross correlation coefficient R, where $0 \leq R \leq 1$; determining whether the cross correlation coefficient R is above a predetermined threshold; and when the cross correlation coefficient R is above a predetermined threshold, communicating with the display screen to display the identifying information associated with the speaker profile.

[0004] Also disclosed is a system for identifying a person on a display screen based on voice recognition. In one embodiment, the system comprises a memory configured to store a speaker profile associated with a speaker, the speaker profile including a signature pattern and identifying information associated with the speaker, the signature pattern having been created from a Fast Fourier Transform (FFT) of a first voice sample of the speaker's voice; a microphone

for receiving a second voice sample; a display screen; and a processor in communication with the microphone, the memory, and the display screen. In this embodiment, the processor is configured to perform the following process: convert the second voice sample into a voice pattern using FFT; cross correlate the voice pattern with the signature pattern from the memory; calculate a cross correlation coefficient R associated with the cross correlation of the voice pattern with the signature pattern, where $0 \leq R \leq 1$; determine whether the cross correlation coefficient R is above a predetermined threshold; and when the cross correlation coefficient R is above a predetermined threshold, instruct the display screen to display the identifying information associated with the speaker profile.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] For a more complete understanding of the present invention, reference is made to the following detailed description of an embodiment considered in conjunction with the accompanying drawings, in which:

[0006] FIG. 1 is a high-level flow chart showing a voice sample collection method accordance with an embodiment of the present invention;

[0007] FIG. 2a is an FFT image of a sample of a first speaker's voice to be used in connection with a profile of the first speaker;

[0008] FIG. 2b is an FFT image of a sample of a second speaker's voice to be used in connection with a profile of the second speaker;

[0009] FIG. 2c is an FFT image of a sample of a third speaker's voice to be used in connection with a profile of the third speaker;

[0010] FIG. 2d is an FFT image of a sample of a fourth speaker's voice to be used in connection with a profile of the fourth speaker;

[0011] FIG. 3 is an FFT image of the captured audio of a conversation where multiple speakers spoke;

[0012] FIG. 4 is a schematic view of a voice recognition system cross correlating the FFT image of FIG. 2d against the FFT image of FIG. 3;

[0013] FIG. 5a is a schematic view of a display screen showing a first person's identifying information; and

[0014] FIG. 5b is a schematic view of a display screen showing a second person's identifying information.

DETAILED DESCRIPTION

[0015] The present disclosure generally relates to a voice-recognition system and method where a person speaking through an audio communication medium is identified to the receiving party by populating a profile unique to the speaker on a display screen on the receiving party's end. The system does so by sampling a speaker's voice while he or she is speaking and, through spectral analysis and the use of sliding Fast Fourier Transforms ("FFTs"), cross-correlates each sample to a number of voice patterns stored in the system, where each voice pattern is associated with a profile. When the cross-correlation of the speaker's voice against the stored voice patterns results in a match with one of the stored voice patterns, the system sends a signal to an electronic device with a display screen on the listening party's end, the signal indicating that the profile associated with the matched voice pattern should appear on the display screen. This profile will identify who the speaker is to the listening party so as to remove any confusion on the listening party's part as to who may be speaking at that given time.

[0016] It should be understood that the elements shown in the figures can be implemented in various forms of hardware, software or combinations thereof. Preferably, these elements are implemented in a combination of hardware and software on one or more appropriately programmed general-purpose devices, which may include a processor, memory and input/output interfaces. Other elements can be implemented through the use of specifically-purposed devices, such as microphones, audio speakers, and electronic display screens.

[0017] Turning now to FIG. 1, a high-level flow chart of voice sample collection method for teleconferencing 100 in accordance with an embodiment of the present invention is shown. As is typical during a teleconference call, this embodiment involves at least two speaking persons. In one embodiment, multiple persons participate in the call by using their own telephone receivers to call in to a teleconferencing service. In another embodiment, one set of multiple persons are using one telephone receiver to speak to a second set of multiple persons on a second telephone receiver.

[0018] At the beginning of the conference call 102, each person on the call provides a sample of his or her voice to the system (steps 104(a)-(n)) before connecting to the call. In one

embodiment, this is done at a stage where each person is asked to state his or her name before being connected. Upon receiving a sample of each person's voice, the system performs a spectral analysis of the voice sample via a Fast Fourier Transform ("FFT") to create a spectral pattern showing the dominant frequencies in the person's voice (steps 106(a)-(n)). Each spectral pattern is then stored in a memory as a signature pattern for each individual speaker, respectively (steps 108(a)-(n)). Each signature pattern is then associated with an individual unique speaker profile (steps 110(a)-(n)). The signature patterns and the speaker profiles are then stored in the system's memory (step 112) and the speakers on the call are then connected to each other to conduct the conference call (step 114).

[0019] FIGs. 2a-d show four examples of speaker profiles 200a-d that were collected from four individual speakers A-D, respectively, using the embodiment illustrated in FIG. 1. As can be seen, each of speaker profiles 200a-d includes a signature pattern (e.g., one of signature patterns 202a-d, respectively) which illustrates three dominant vocal frequencies that characterize the voice of the individual speaker who gave the voice sample. For example, signature pattern 202a for individual speaker A shows three dominant frequencies of 1 kHz, 500 Hz, and 250 Hz that characterize the voice of individual speaker A, while signature pattern 202b for individual speaker B shows three dominant frequencies of 800 Hz, 600 Hz, and 200 Hz that characterize the voice of individual speaker B. These signature patterns 202a-d are used as bases upon which voice identification can take place during a conference call between individual speakers A-D. This voice identification method is discussed further below.

[0020] Still referring to FIGs. 2a-d, in addition to the signature patterns 202a-d, each of the speaker profiles 200a-d also includes speaker identification information (not shown) particular to the speaker associated with the speaker profile. This speaker identification information is used to inform others on the conference call who is speaking once the voice identification method has identified a particular speaker as speaking at a given time during the call. In one embodiment, the identification information includes the speaker's name. In other embodiments, the identification information can include, but is not limited to, the speaker's name and a picture of the speaker's face and the like.

[0021] During a conference call, the system takes samples of the conversation in real time and converts each sample as it is taken into an FFT image and/or pattern highlighting the dominant frequencies in the sample. FIG. 3 illustrates an FFT profile 300 of a captured

conversation during such a conference call. The FFT profile 300 exhibits where windowed filters captured samples of the conversation at a-q instances in time. The a-q instances represent the intervals of time at which a voice sample was captured during the conversation. In one embodiment, the intervals are 0.5 seconds long. In another embodiment, the intervals are 0.25 seconds long. At each of the a-q instances in time, a spectral pattern is given representing the dominant frequencies, if any at all, that were present during the conversation at that given instance. In the example shown in FIG. 3, the spectral graph 300 shows nine instances where a voice sample was taken and converted via FFT into a voice pattern (e.g., voice patterns 302-318).

[0022] FIG. 4 illustrates how speakers are identified from voice samples taken during the conversation. During the sampled conversation, each of the signature patterns 202a-d found in the speaker profiles 200a-d is slid across the FFT profile of the conversation to determine whether the person associated with the selected signature pattern is speaking. At each instance of time *a-q*, the system then cross correlates each of the signature patterns 202a-d against the converted voice sample captured at that instance of time. Each cross correlation produces a correlation coefficient *R* having a value between 0 and 1, where 1 indicates a perfect match and a 0 indicates no match. When the correlation efficient R produces a value above a predetermined threshold, the system sends a signal indicating that the individual speaker associated with the selected signature pattern is currently speaking.

[0023] In the example shown in FIG. 4, signature pattern 202d from speaker profile 200d is slid across the conversation's FFT profile 300 and cross correlated with the captured sample at each of the a-q instances in time during the conversation. At instances *a* and *b*, the cross correlation coefficient *R* will be 0, since no one was speaking during those instances of time. However, at instance *c*, the cross correlation of voice pattern 302 with signature pattern 202d results in an *R* value of 1, indicating that speaker D is speaking at that instance in time. This is also the case for instances *o* and *q*, where correlations with voice patterns 316 and 318 also result in an *R* value of 1. At instances of time where a voice sample produces a voice pattern that is not the same as that of speaker D (such as voice patterns 304, 306, 308, 310, 312, and 314), the cross correlation with signature pattern 202d reveals an R value between 0 and 1. In such instances, a predetermined threshold for *R* is set high enough such that other voice patterns correlated with

signature pattern 202d do not cause the system to signal false positive identifications. In one embodiment, a predetermined threshold for *R* is 0.8.

[0024] Once an R value associated with a cross correlation of a particular signature pattern reaches or exceeds a predetermined threshold, the systems sends a signal to a display screen to display the identification information from the speaker profile associated with the particular signature pattern. The display screen then displays the identification information to inform those listening to the call who is speaking at that instance in time. FIGS. 5a and 5b illustrate examples of a display screen 500 displaying identification information 502a, 502b upon receiving a signal from the system to do so. To use the examples shown in FIGS. 2a-d, 3, and 4, the display screen 500 displays the identification information 502a from speaker profile 200a when it receives a signal indicating a match with signature pattern 202a and displays the identification information 502b from speaker profile 200b when it receives a signal indicating a match with signature pattern 202b. This allows those on the call within view of the display screen 500 to see that speakers A and B are speaking on the call when their respective identification information 502a, 502b appears on the display screen 500.

[0025] The system can also send identification information to multiple devices such as mobile devices including, but not limited to, laptops, tablets and/or mobile phones and the like. This can be accomplished via wired (e.g., Ethernet cables, etc.) and/or wireless means (e.g., Bluetooth, WiFi, etc.). This allows easier access to the identification information. It also permits each user to customize how the information is displayed on their device. For example, when a senior person with high authority speaks, the information can be shown in red or bordered by a red border to gain the user's attention more quickly.

[0026] Many variations can be made to the systems and methods discussed above. For example, in one embodiment, as the system performs cross-correlations of the voice patterns in the conversation against the stored signature patterns, the system can also take the voice patterns which have been identified as a match to one of the signature patterns and average them together with the identified signature pattern as the conversation progresses. This allows the system to learn the characteristics of each person's voice as the conversation progresses and can potentially lead to more accurate identifications of the various speakers during the call.

[0027] The various embodiments disclosed herein can be implemented as hardware, firmware, software, or any combination thereof. Moreover, the software is preferably

implemented as an application program tangibly embodied on a program storage unit or computer readable medium. The application program may be uploaded to, and executed by, a machine comprising any suitable architecture. Preferably, the machine is implemented on a computer platform having hardware such as one or more central processing units ("CPUs"), a memory, and input/output interfaces. The computer platform may also include an operating system and microinstruction code. The various processes and functions described herein may be either part of the microinstruction code or part of the application program, or any combination thereof, which may be executed by a CPU, whether or not such computer or processor is explicitly shown. In addition, various other peripheral units may be connected to the computer platform such as an additional data storage unit and a printing unit.

[0028] All examples and conditional language recited herein are intended for pedagogical purposes to aid the reader in understanding the principles of the invention and the concepts contributed by the inventor to furthering the art, and are to be construed as being without limitation to such specifically recited examples and conditions. Moreover, all statements herein reciting principles, aspects, and embodiments of the invention, as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended that such equivalents include both currently known equivalents as well as equivalents developed in the future, i.e., any elements developed that perform the same function, regardless of structure.

[0029] It will be understood that the embodiments described herein are merely exemplary and that a person skilled in the art may make many variations and modifications without departing from the spirit and scope of the invention. All such variations and modifications are intended to be included within the scope of the invention as defined in the appended claims.

CLAIMS:

1. A voice recognition method for a communication system, the method comprising:

obtaining a speaker profile associated with a speaker, the speaker profile including a signature pattern and identifying information associated with the speaker, the signature pattern created from a spectral analysis of a first voice sample of the speaker's voice;

receiving a second voice sample;

converting the second voice sample into a voice pattern using spectral analysis;

cross correlating the voice pattern with the signature pattern, resulting in a cross correlation coefficient R, where $0 \leq R \leq 1$; and

sending the identifying information associated with the speaker profile to at least one device when the cross correlation coefficient R is above a threshold.

2. The voice recognition method according to claim 1, wherein the identifying information includes the speaker's name.

3. The voice recognition method according to claim 1, wherein the identifying information includes an image associated with the speaker.

4. The voice recognition method according to claim 1, wherein the method is repeated X times over X instances of time, and wherein the second voice sample is a new voice sample at each of the X instances of time, wherein X is an integer.

5. The voice recognition method according to claim 1, wherein obtaining the speaker profile includes receiving the first voice sample from the speaker, converting the first voice sample into the signature pattern using FFT, and associating the identifying information with the signature pattern to create the speaker profile.

6. The voice recognition method according to claim 1, wherein a threshold is 0.8.

7. A system for identifying a person based on voice recognition, the system comprising:

a memory configured to store a speaker profile associated with a speaker, the speaker profile including a signature pattern and identifying information associated with the speaker, the signature pattern created from a spectral analysis of a first voice sample of the speaker's voice;

a microphone for receiving a second voice sample; and

a processor in communication with the microphone, the memory, and the display device, the processor configured to perform the following process:

convert the second voice sample into a voice pattern using spectral analysis;

cross correlate the voice pattern with the signature pattern from the memory;

calculate a cross correlation coefficient R associated with the cross correlation of the voice pattern with the signature pattern, where $0 \leq R \leq 1$; and

send the identifying information associated with the speaker profile to at least one device when the cross correlation coefficient R is above a threshold,.

8. The system according to claim 7, wherein the identifying information includes the speaker's name.

9. The system according to claim 7, wherein the identifying information includes an image associated with the speaker.

10. The system according to claim 7, wherein the processor is further configured to repeat the process X times over X instances of time, and wherein the second voice sample is a new voice sample at each of the X instances of time, wherein X is an integer.

11. The system according to claim 7, wherein a threshold is 0.8.

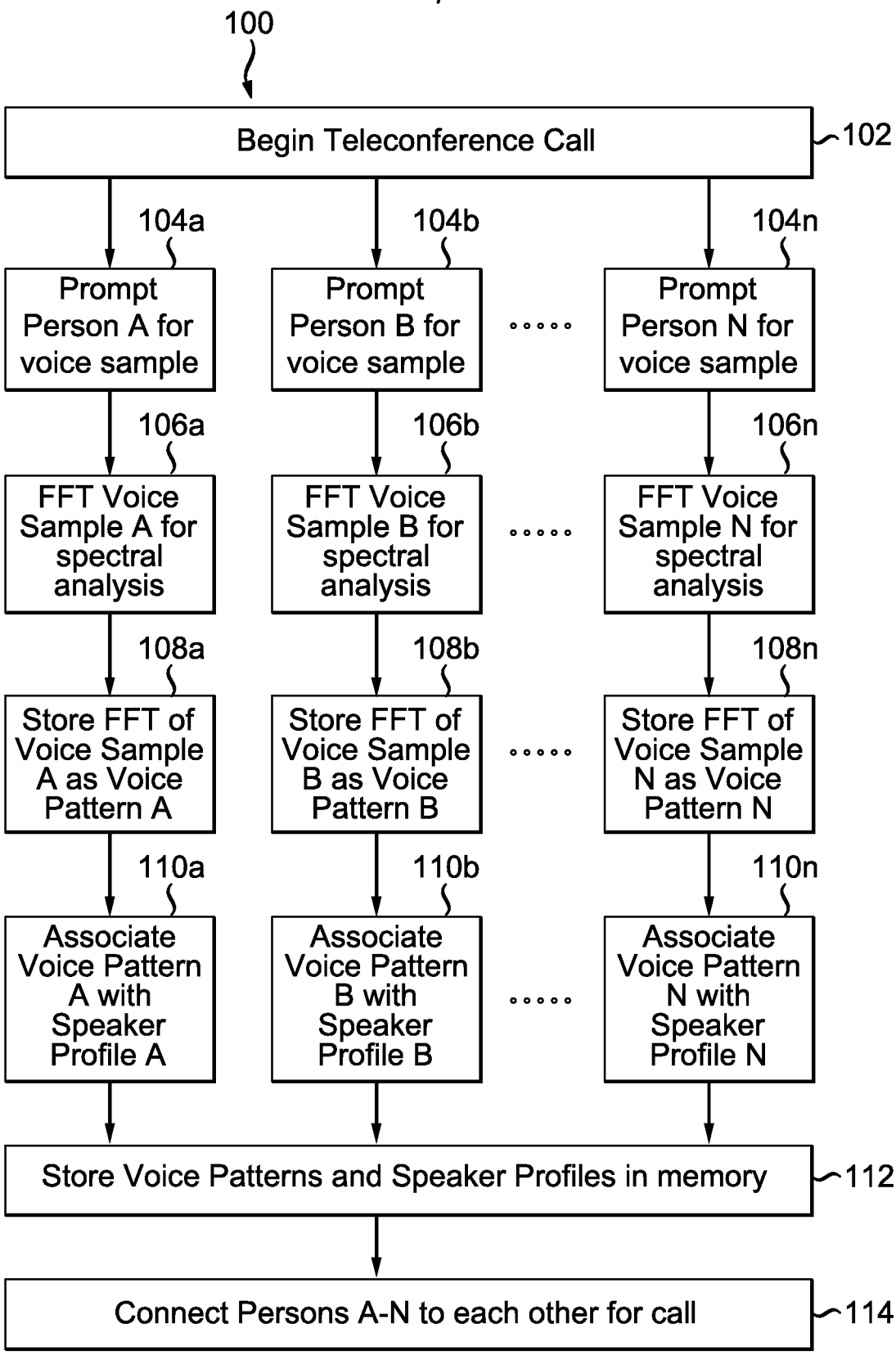

100
Begin Teleconference Call
102
104a
104b
104n
Prompt Person A for voice sample
Prompt Person B for voice sample
Prompt Person N for voice sample
106a
106b
106n
FFT Voice Sample A for spectral analysis
FFT Voice Sample B for spectral analysis
FFT Voice Sample N for spectral analysis
108a
108b
108n
Store FFT of Voice Sample A as Voice Pattern A
Store FFT of Voice Sample B as Voice Pattern B
Store FFT of Voice Sample N as Voice Pattern N
110a
110b
110n
Associate Voice Pattern A with Speaker Profile A
Associate Voice Pattern B with Speaker Profile B
Associate Voice Pattern N with Speaker Profile N
Store Voice Patterns and Speaker Profiles in memory
112
Connect Persons A-N to each other for call
114

FIG. 1

FIG. 2A

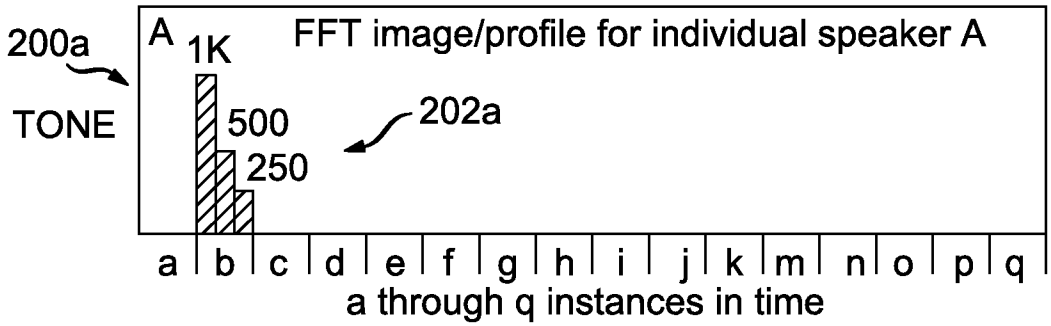


FIG. 2B

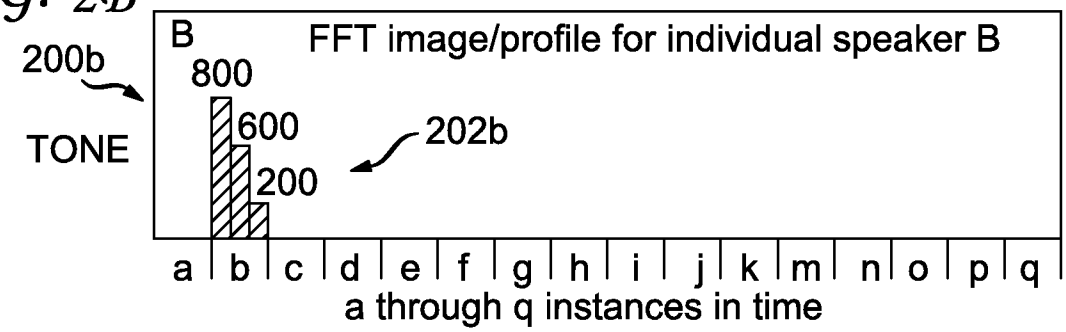


FIG. 2C

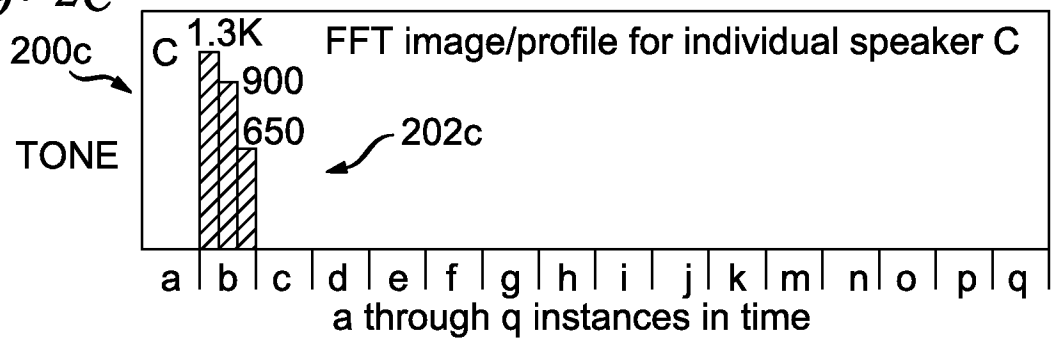
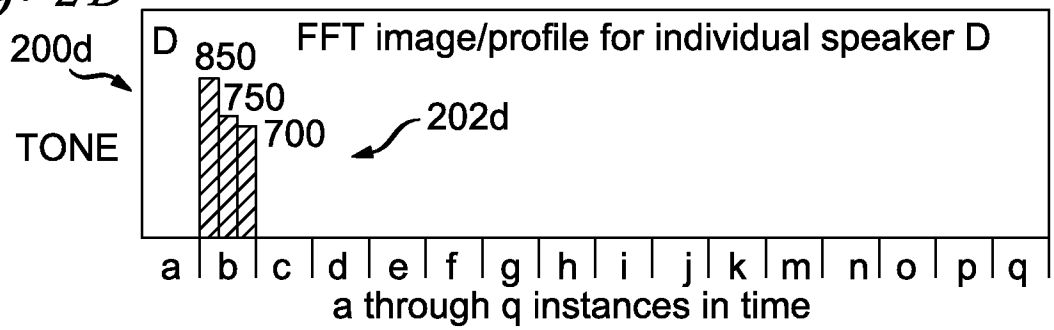


FIG. 2D



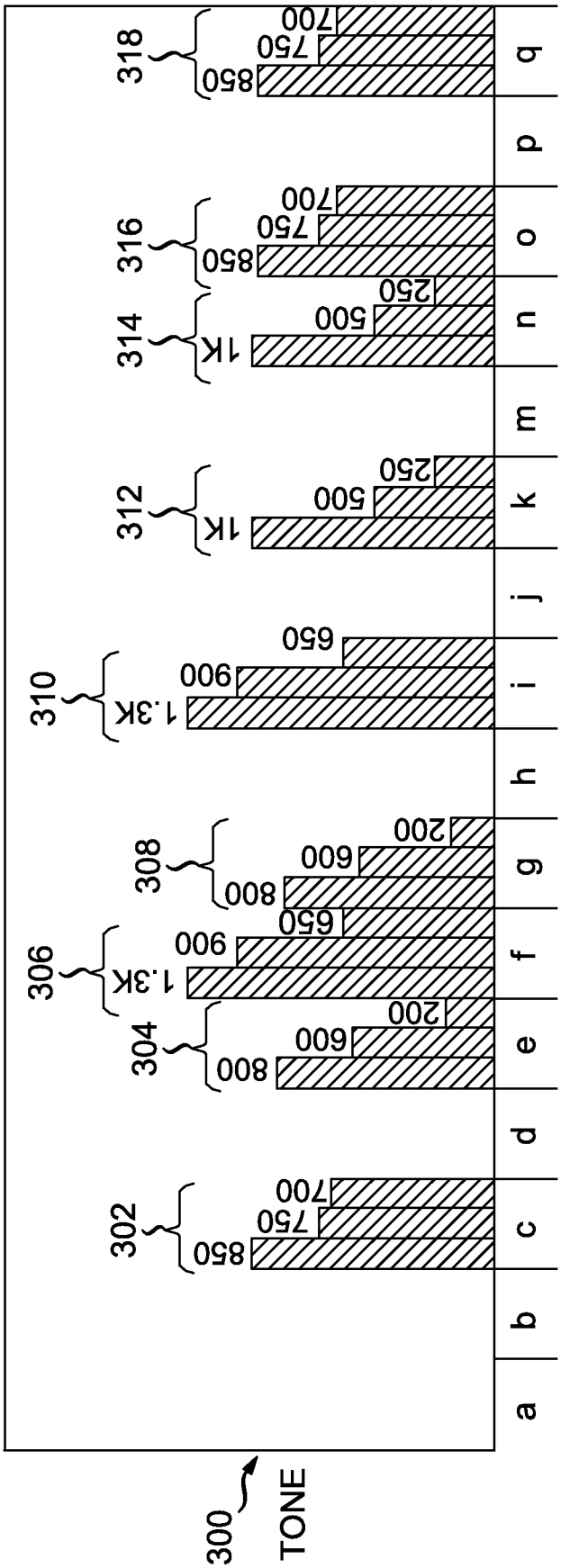

300
TONE
a through q instances in time
a
b
c
d
e
f
g
h
i
j
k
m
n
o
p
q
302
850
750
700
304
800
600
200
306
1.3K
900
650
308
800
600
200
310
1.3K
900
650
312
1K
500
250
314
1K
500
250
316
850
750
700
318
850
750
700

FIG. 3

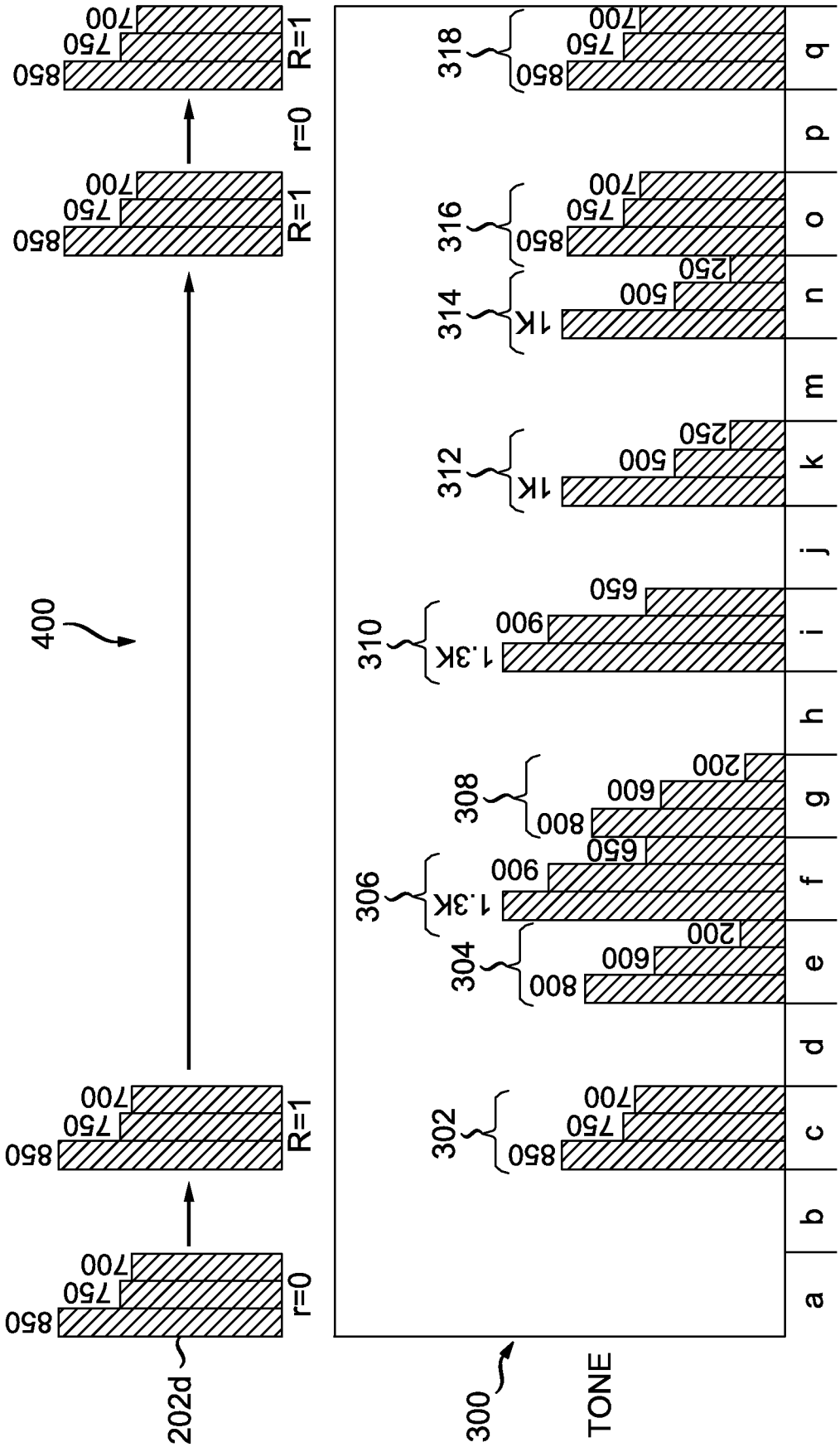

202d
300
TONE
400
r=0
R=1
850
750
700
302
304
306
308
310
312
314
316
318
1K
1.3K
500
250
650
900
600
200
800
a through q instances in time
a
b
c
d
e
f
g
h
i
j
k
m
n
o
p
q

FIG. 4

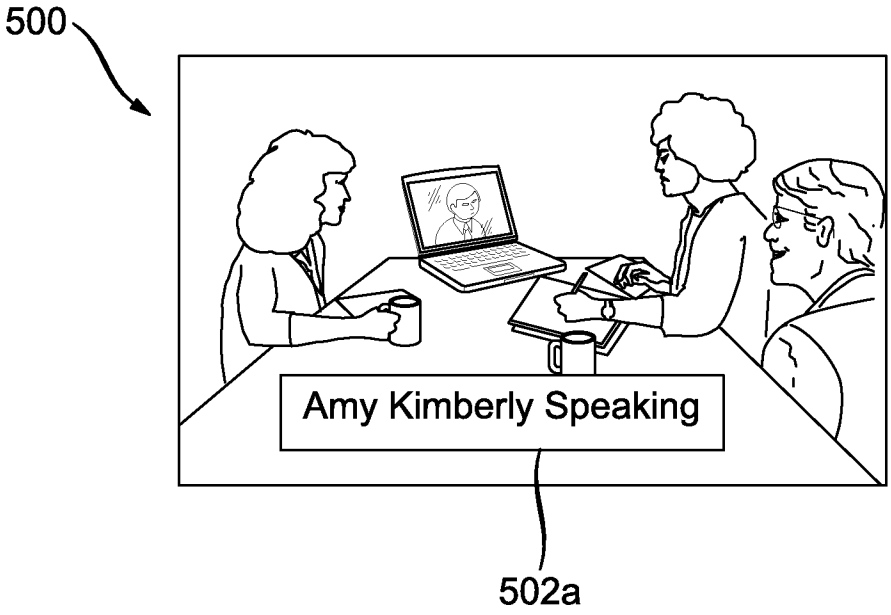


FIG. 5A

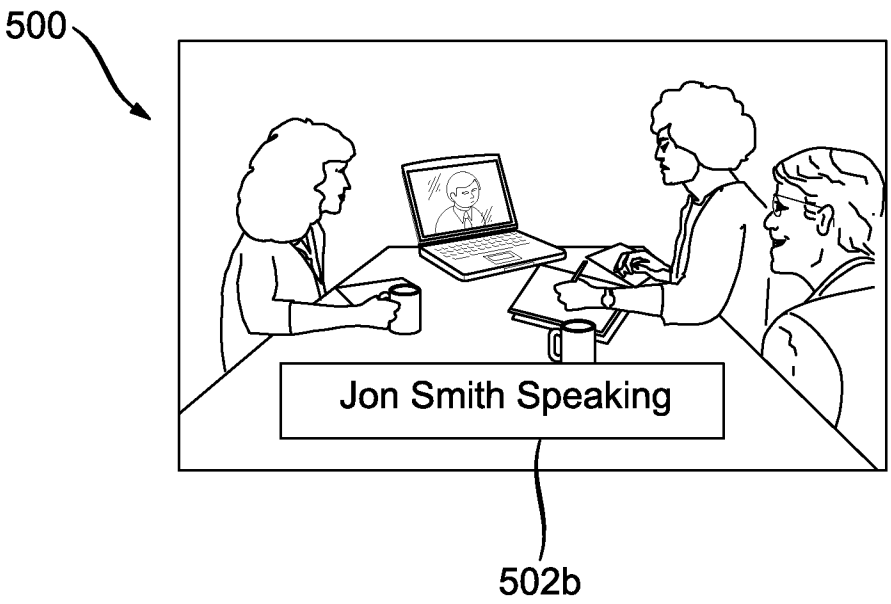


FIG. 5B

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/044413

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L17/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, COMPENDEX, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 718 306 B1 (SATOH KATSUHIKO [JP] ET AL) 6 April 2004 (2004-04-06) abstract column 6, line 36 - column 9, line 34 -----	1-11
X	Rakesh D R ET AL: "SPEAKER RECOGNITION AND AUTHENTICATION", International Journal of Computer Science and Mobile Computing, 30 May 2013 (2013-05-30), pages 402-407, XP055114301, Retrieved from the Internet: URL:http://ijcsmc.com/docs/papers/May2013/V2I52013136.pdf the whole document ----- -/--	1-11

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 April 2014	28/04/2014

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Deltorn, Jean-Marc

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/044413

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	DHANANJAYA N ET AL: "Correlation-Based Similarity Between Signals for Speaker Verification with Limited Amount of Speech Data", 1 January 2006 (2006-01-01), MULTIMEDIA CONTENT REPRESENTATION, CLASSIFICATION AND SECURITY LECTURE NOTES IN COMPUTER SCIENCE;;LNCS, SPRINGER, BERLIN, DE, PAGE(S) 17 - 25, XP019040403, ISBN: 978-3-540-39392-4 the whole document --- ZALEWSKI J ET AL: "Cross Correlation of Long-Term Speech Spectra as a Speaker Identification Technique", ACUSTICA UNITED WITH ACTA ACUSTICA, S. HIRZEL VERLAG, STUTTGART, DE , vol. 34, no. 1 1 November 1975 (1975-11-01), pages 20-24, XP008168845, ISSN: 1610-1928 Retrieved from the Internet: URL:http://www.ingentaconnect.com/content/dav/aaua/1975/00000034/00000001/art00005 the whole document --- US 6 931 113 B2 (ORTEL WILLIAM CHARLES GORMLEY [US]) 16 August 2005 (2005-08-16) the whole document ---	1-11 1-11 1-11

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2013/044413

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6718306 B1	06-04-2004	CN 1327575 A	19-12-2001
		DE 60014031 D1	28-10-2004
		DE 60014031 T2	20-01-2005
		EP 1141943 A1	10-10-2001
		JP 2001117579 A	27-04-2001
		KR 100426351 B1	08-04-2004
		US 6718306 B1	06-04-2004
		WO 0129824 A1	26-04-2001
US 6931113 B2	16-08-2005	US 2004091086 A1	13-05-2004
		US 2005207554 A1	22-09-2005
		US 2011069140 A1	24-03-2011